EXHIBIT 125

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

IN RE MATTER OF:	)	
RICHARD KADREY, et al.,	)	
Plaintiff,	)	
vs.	)	C.A. NO.:
META PLATFORMS, INC.,	)	3:23-cv-03417-VC
Defendant.	)	
	)	

VIDEOTAPED DEPOSITION OF YANN LeCUN, Ph.D.
Palo Alto, California
Thursday, November 21, 2024
** HIGHLY CONFIDENTIAL - ATTORNEYS EYES ONLY **
UNDER THE PROTECTIVE ORDER

Stenographically Reported by:
HEATHER J. BAUTISTA, CSR, CRR, RPR, CLR
Realtime Systems Administrator
California CSR License #11600
Oregon CSR License #21-0005
Washington License #21009491
Nevada CCR License #980
Texas CSR License #10725

DIGITAL EVIDENCE GROUP
1730 M. Street, NW, Suite 812
Washington, D.C. 20036
(202) 232-0646

If you have an understanding of the question independent of that, by means you can provide it. 10:27:57 10:28:00

MS. GEMAN: That question was plainly not one that would beget that admonition, so I would respectfully request that you keep your objections to form. 10:28:04 10:28:05 10:28:10 10:28:12

MR. WEINSTEIN: I -- I understand that. I understand you weren't meaning to get that, but the scope of your question is pretty broad and -- 10:28:12 10:28:13 10:28:15

Q. (By Ms. Geman) Do you need to hear the question again? 10:28:18 10:28:19

A. No. I think there are two aspects. The first one is a completely practical one which is that those systems are -- are trained -- and the answer is different depending on which version of -- 10:28:20 10:28:22 10:28:29 10:28:32

(Stenographer clarification.) 10:28:36

THE WITNESS: -- LLaMA which is the -- 10:28:37

Q. (By Ms. Geman) So I will -- I can make the question more precise. 10:28:39 10:28:40

A. Yes. 10:28:41

Q. Do you agree with Meta's decision to fail to disclose the training sources for LLaMA 2, 3, and 10:28:42 10:28:43

4? 10:28:50

A. So 4 doesn't exist yet. 10:28:53

Q. I understand that. 10:28:55

A. So there's two things you can list; right? 10:28:56

So you can list the -- the set of datasets that we used to train the system which was done for LLaMA 1. It was not done for LLaMA 2 and 3. 10:29:01 10:29:10 10:29:13

You can also just redistribute the data and that wasn't done for any of the models. For LLaMA 1, all those sources were publicly downloadable from the open Internet. 10:29:16 10:29:19 10:29:24 10:29:28

For LLaMA 3, that wouldn't be possible because some of the data comes from -- it's internal data essentially from users of Meta's services, and so it cannot really be distributed, although that content is public. I mean only constituted of public posts and pictures from -- from users of Meta's platforms. So that would not be distributed. 10:29:30 10:29:32 10:29:37 10:29:44 10:29:47 10:29:52 10:29:57

There's also the fact that it's not really useful for that data to be distributed in the first place because the reason for open source -- 10:30:05 10:30:09 10:30:12

(Stenographer clarification.) 10:30:20

THE WITNESS: -- open sourcing a foundation model, such as LLaMA, is -- is to allow people to use this model without having to train it from scratch which is insanely expensive. So giving the data would not help people particularly. 10:30:22 10:30:25 10:30:30 10:30:32 10:30:41

And there's a lot of work that goes into curating a training dataset. Basically filtering it in such a way that it only contains high-quality data, and that is not viewed as something that should be publicly revealed for various reasons. 10:30:45 10:30:52 10:30:59 10:31:03 10:31:09

Q. (By Ms. Geman) So my question is: Do you agree in Meta's decision to fail to disclose the training sources for LLaMA 2 and 3? 10:31:13 10:31:15 10:31:18

A. So that was a decision by our legal department and I have no input. 10:31:21 10:31:24

Q. But you're the chief scientist and a very well-regarded scholar and practitioner of AI, so my question is: Do you agree with that decision? 10:31:28 10:31:30 10:31:36

MR. WEINSTEIN: Okay. 10:31:38

You can state your personal opinion. Her question isn't asking for it, but don't reveal the communications you had with counsel. 10:31:38 10:31:40 10:31:43

plural or not, and so typically a token represents about three-quarter of a word on average. 12:41:44 12:41:47

And then there are -- 12:41:50

(Stenographer clarification.) 12:41:55

THE WITNESS: -- agglutinative languages such as German where a word can be constructed by sticking together many words, right, so you have to break that word into individual components. So that's what a token is, really; it's kind of a unit of language that could be smaller than the word. 12:41:55 12:41:57 12:42:02 12:42:04 12:42:08 12:42:12

Q. (By Ms. Geman) So let's just say that was the only sentence the LLM trained on it, it would have a set of tokens for that sentence, right -- 12:42:16 12:42:17 12:42:21

A. Yeah. 12:42:23

Q. -- "a sheep, a wolf and a cabbage lived on an island." 12:42:24 12:42:28

Would -- if you were to give -- and every single time it had -- every single time it trained on that sentence, let's say, it would have that same group of tokens; right? Okay. 12:42:29 12:42:33 12:42:36 12:42:39

And if you were to sort of show the LLM that group of tokens, it could spit out that 12:42:43 12:42:47

sentence; correct? 12:42:51

MR. WEINSTEIN: Object to form. 12:42:54

THE WITNESS: That's a tricky -- tricky 12:42:57

answer to that question. 12:43:02

Q. (By Ms. Geman) That sentence -- 12:43:07

A. No, I can answer the question, but -- 12:43:07

Q. Okay. 12:43:08

A. Okay. 12:43:10

Q. Sorry, we're noticing it was an hour and 12:43:12

counsel was looking at me, so that's what were 12:43:14

noticing. 12:43:15

But please -- please go ahead and then we 12:43:15

can do our lunch break. 12:43:17

A. Okay. 12:43:19

The same way that if I give you a short 12:43:19

poem and I keep repeating it to you multiple times, 12:43:22

you will learn that poem by -- by heart. But you're 12:43:28

going to be able to regurgitate that poem, if that's 12:43:33

the only thing you hear for two days; right? 12:43:36

But if the space of those two days you read 12:43:39

two novels, you're not going to be able to 12:43:42

regurgitate all the words in those novels. 12:43:46

It's exactly the same thing for LLMs. If you train them on a small amount of data and it's a big LLM with a big capacity, memory capacity, and you train it -- you train it on a small amount of text, it's probably going to be able to regurgitate that, the text you've been training it on, with some degree of fidelity. 12:43:48 12:43:50 12:43:54 12:43:56 12:43:58 12:44:02 12:44:05

But if you train it on a lot of data, it just does not have the memory capacity to -- to store the content and -- and therefore it's not going to be able to regurgitate any of those texts to any significant extent unless within the training set, a particular short segment of text appears thousands of times. 12:44:08 12:44:11 12:44:14 12:44:17 12:44:21 12:44:26 12:44:30

So a famous quote, for example, may appear thousands of times in a training set and the LLM obviously would be able to regurgitate that quote because, you know, it appears so many times and it's probably part of common language if it appears so many times, but it's not going to be able to regurgitate the whole novel or even a chapter or even a paragraph probably. 12:44:31 12:44:35 12:44:38 12:44:41 12:44:44 12:44:47 12:44:50 12:44:53

Q. If a large LLM trained on only one book, it could regurgitate -- it could regurgitate that book; isn't that true? 12:44:55 12:44:59 12:45:02

A. The answer to that question depends on the length of the book and the size of the LLM. So if you train a small language model on a single book, it's not going to be able to regurgitate it. 12:45:07 12:45:09 12:45:13 12:45:16

Q. LLaMA could; correct? 12:45:19

A. Well, there's sizes of LLaMA. LLaMA 3 has -- 12:45:21 12:45:25

(Stenographer clarification.) 12:45:27

THE WITNESS: -- several versions of various sizes. The smallest one is 1 billion parameters. And then there is a 7 billion, 70 billion and a 405 billion. 12:45:27 12:45:28 12:45:33 12:45:34

Q. (By Ms. Geman) Are you -- are you disputing that LLaMA 3 could regurgitate a book if that's all it was trained on? 12:45:36 12:45:39 12:45:44

A. If it was trained on a single book, not too long, it might be able to store the entire book in its memory and regurgitate it relatively accurately, but nobody would want an LLM of this type. It would 12:45:45 12:45:50 12:45:55 12:46:00

be an extremely inefficient way of storing a book and inaccurate and also expensive. 12:46:03 12:46:08

MS. GEMAN: Okay. 12:46:11

We can take a break. 12:46:11

THE WITNESS: Okay. 12:46:13

THE VIDEOGRAPHER: We're going to go off the record. The time is 12:46 p.m. 12:46:14 12:46:15

(Lunch recess 12:46 p.m. to 1:42 p.m.) 12:46:26

THE VIDEOGRAPHER: We are back on the record. The time is 1:42 p.m. 12:46:26 13:42:19

Please proceed. 13:42:21

Q. (By Ms. Geman) Hi, Yann. Did you have a good lunch? 13:42:25 13:42:26

A. Same as yours, I'm afraid. 13:42:27

Q. Okay. 13:42:34

I notice you didn't answer if it was a good lunch. 13:42:34 13:42:37

What is memorization -- what is LLM memorization? 13:42:37 13:42:42

A. So it's something that does not really exist in the sense that -- I think as we alluded to earlier, when you train an LLM on a large 13:42:43 13:42:44 13:42:47

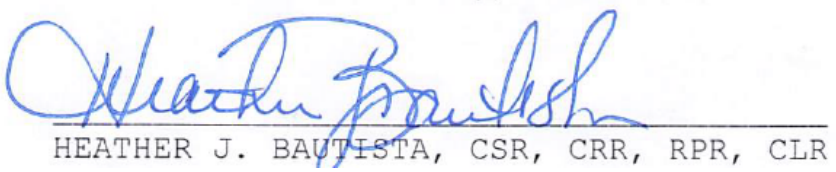
I, HEATHER J. BAUTISTA, CSR No. 11600, Certified Shorthand Reporter, certify:

That the foregoing proceedings were taken before me at the time and place therein set forth, at which time the witness declared under penalty of perjury; that the testimony of the witness and all objections made at the time of the examination were recorded stenographically by me and were thereafter transcribed under my direction and supervision; that the foregoing is a full, true, and correct transcript of my shorthand notes so taken and of the testimony so given;

(XX) Reading and signing was not requested/offered.

I further certify that I am not financially interested in the action, and I am not a relative or employee of any attorney of the parties, nor of any of the parties.

I declare under penalty of perjury under the laws of California that the foregoing is true and correct. Dated: November 26, 2024



HEATHER J. BAUTISTA, CSR, CRR, RPR, CLR